Concentration Inequalities for Statistical Learning

Huiming Zhang

Postdoctoral Research Fellow[1] and Research Associate[2]

1.Department of Mathematics, Faculty of Science and Technology, University of Macau;
2. Zhuhai UM Science & Technology Research Institute.

Email: huimingzhang@um.edu.mo

Deep Learning Lecture Notes Chapter 17: Concentration of measure
`https://mjt.cs.illinois.edu/dlt/index.pdf`

Presented in CUHK(SZ) (Online)

April 20, 2021

Overview

Commun. Math. Res.
doi: 10.4208/cmr.2020-0041

Vol. 37, No. 1, pp. 1-85
February 2021

REVIEW ARTICLE

Concentration Inequalities for Statistical Inference

Huiming Zhang[1,3,4] and Song Xi Chen[1,2,3,*]

[1] *School of Mathematical Sciences, Peking University, Beijing 100871, P.R. China.*
[2] *Guanghua School of Management, Peking University, Beijing 100871, P.R. China.*
[3] *Center for Statistical Sciences, Peking University, Beijing 100871, P.R. China.*
[4] *Department of Mathematics, Faculty of Science and Technology, University of Macau, P.R. China.*

Received 4 November 2020; Accepted 12 December 2020

Abstract. This paper gives a review of concentration inequalities which are widely employed in non-asymptotical analyses of mathematical statistics in a wide range of settings, from distribution-free to distribution-dependent, from sub-Gaussian to sub-exponential, sub-Gamma, and sub-Weibull random variables, and from the mean to the maximum concentration. This review provides results in these settings with some fresh new results. Given the increasing popularity of high-dimensional data and inference, results in the context of high-dimensional linear and Poisson regressions are also provided. We aim to illustrate the concentration inequalities with known constants and to improve existing bounds with sharper constants.

AMS subject classifications: 60F10, 60G50, 62E17
Key words: Constants-specified inequalities, sub-Weibull random variables, heavy-tailed distributions, high-dimensional estimation and testing, finite-sample theory, random matrices.

1 Introduction

In probability theory and statistical inference, researchers often need to bound the probability of a difference between a random quantity from its target, usually the error bound of estimation. Concentration inequalities (CIs) are tools for attaining such bounds, and play important roles in deriving theoretical results for various inferential situations in statistics and probability. The recent developments in high-dimensional (HD) statistical inference, and statistical and machine learning have generated renewed interests in the CIs, as reflected in [29, 47, 84, 86]. As the CIs are diverse in their forms and the underlying distributional requirements, and are scattered around in references, there is an increasing need for a review which collects existing results together with some new results (sharper and constants-specified CIs) from the authors for researchers and graduate students working in statistics and probability. This motivates the writing of this review.

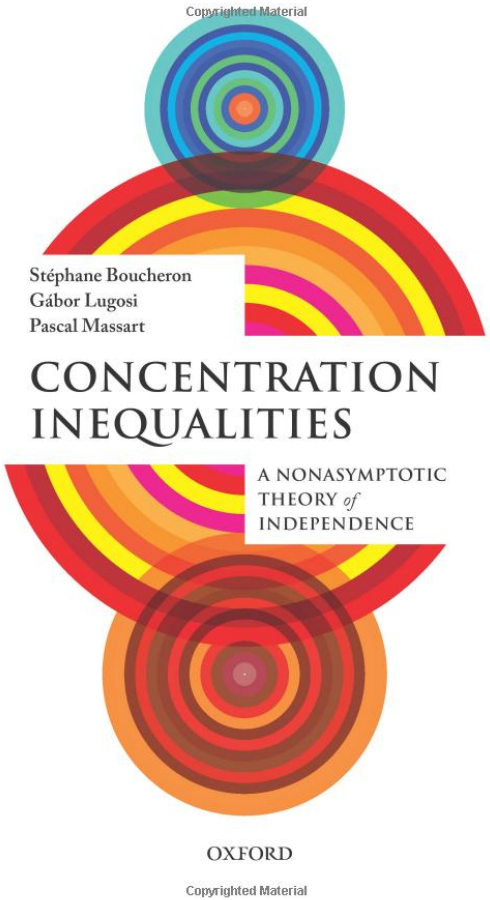
CIs enable us to obtain non-asymptotic results for estimating, constructing confidence intervals, and doing hypothesis testing with a high-probability guarantee. For example, the first-order optimized condition for HD linear regressions should be held with a high probability to guarantee the well-behavior of the estimator. The concentration inequality for error distributions is to ensure the concentration from first-order optimized conditions to the estimator. Our review focuses on four types of CIs:

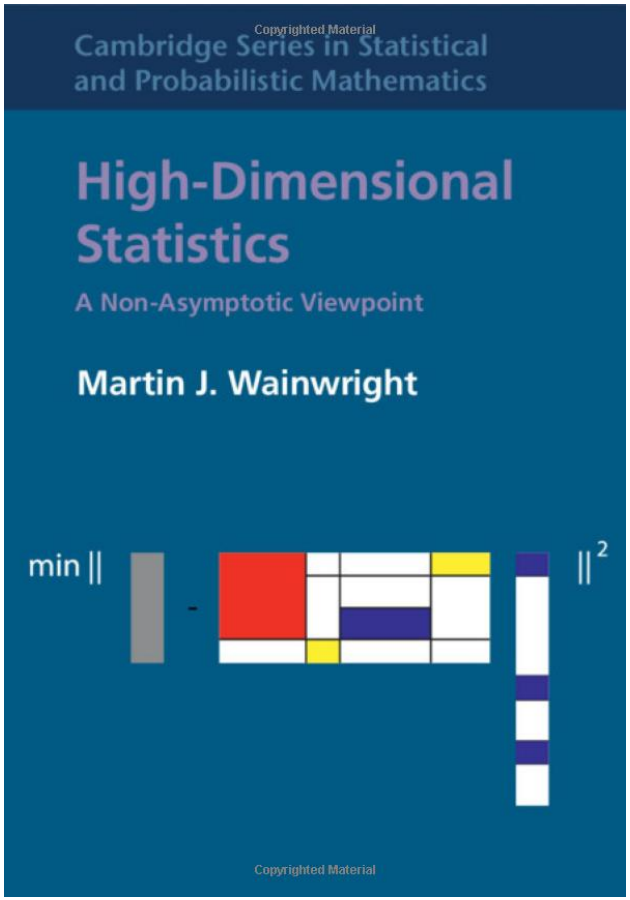
$$P(Z_n > \mathrm{E}Z_n + t), \quad P(Z_n < \mathrm{E}Z_n - t), \quad P(|Z_n - \mathrm{E}Z_n| > t), \quad \mathrm{E}(\max_{i=1,\dots,n} |X_i|),$$

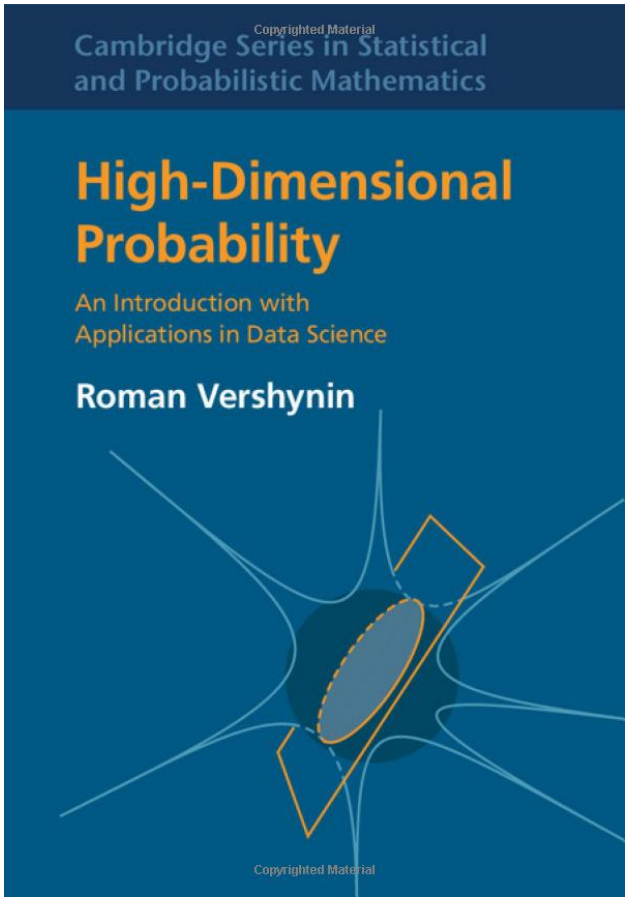
where $Z_n := f(X_1, \cdots, X_n)$ and $X_1, \cdots, X_n$ are random variables. We present two types of CIs: distribution-free and distribution-dependent. Distribution free CIs are free of distribution assumptions, while the distribution-dependent CIs are based on exponential moment conditions reflecting the tail property for the particular class of distributions. Concentration phenomenons for a sum of sub-Weibull random variables will lead to a mixture of two tails: sub-Gaussian for small deviations and sub-Weibull for large deviations from the mean, and it is closely related to Strong Law of Large Numbers, Central Limit Theorem, and Law of the Iterative Logarithm. We provide applications of the CIs to empirical processes and high-dimensional data settings. The latter includes the linear and Poisson regression with a diverging number of covariates. We organize the materials in the forms of lemmas, corollaries, propositions, and theorems. Lemmas and corollaries are on existing results usually without proof except for a few fun-

Figure: **Zhang, H.**, & Chen, S.X. (2021). Concentration Inequalities for Statistical Inference. Communications in Mathematical Research. 37(1), 1-85.

Three books


Copyrighted Material
Stéphane Boucheron
Gábor Lugosi
Pascal Massart
CONCENTRATION
INEQUALITIES
A NONASYMPTOTIC
THEORY of
INDEPENDENCE
OXFORD
Copyrighted Material


Copyrighted Material
Cambridge Series in Statistical
and Probabilistic Mathematics
High-Dimensional
Statistics
A Non-Asymptotic Viewpoint
Martin J. Wainwright
min ||
||2
Copyrighted Material


Copyrighted Material
Cambridge Series in Statistical
and Probabilistic Mathematics
High-Dimensional
Probability
An Introduction with
Applications in Data Science
Roman Vershynin
Copyrighted Material

Starting from Confidence Sets in Undergraduate Statistics

- Given that $\{X_i\}_{i=1}^n \overset{\text{i.i.d.}}{\sim} N(\mu_0, 1)$, we have

$$P\left(\mu_0 \in \left[\bar{X} \pm \frac{1.96}{\sqrt{n}}\right]\right) = 95\%.$$

- Without knowing the distribution of $\{X_i\}_{i=1}^n$, we obtain via CLT

$$P\left(\mu_0 \in \left[\bar{X} \pm \frac{1.96}{\sqrt{n}}\right]\right) \to 95\%$$

However, the price is the "asymptotic" validity.

- What if the distribution of $\{X_i\}_{i=1}^n$ is unknown and n is finite?

Nonasymptotic($\forall$ finite n) Confidence Set:

$P(\mu_0 \in [L_n, U_n]) \geq 1 - \delta$ based on concentration of measure.

(No distribution assumption for **densities**, but few **moment conditions**)

Howard, S. R., Ramdas, A., McAuliffe, J., & Sekhon, J. (2021). Time-uniform, nonparametric, nonasymptotic confidence sequences. The Annals of Statistics, 49(2), 1055-1080.

A first thought: concentration inequality

- Concentration inequalities (CI) is to quantify the concentration of measure.
- Usually, CI **quantifies how a r.v. X deviates around its mean** $\mathrm{E}X =: \mu$ by presenting as one-side or two-sided bounds for **the tail prob. of** $X - \mu$:

$$P(X - \mu > t) \text{ or } P(|X - \mu| > t) \leq \text{ some small } \delta.$$

Good news: CI holds for any sample size n. The simplest CIs are:

Lemma 1.1

Chebyshev's Inequality: *Let X be a r.v. with finite expectation $\mathrm{E}X$ and variance $\mathrm{Var}\, X$. Then, for any $a \in \mathbb{R}^+$:*

$$P(|X - \mathrm{E}X| \geq a) \leq \frac{\mathrm{Var}\, X}{a^2}.$$

Markov's inequality: *Let $\varphi(x) : \mathbb{R} \to \mathbb{R}^+$ be any non-decreasing positive function. For a r.v. X, $P(X \geq a) \leq \mathrm{E}[\varphi(X)]\frac{1}{\varphi(a)}$, $\forall\ a \in \mathbb{R}$.*

The Chebyshev's ineq. prescribes a poly. rate of convergence based on $\mathrm{Var}\, X$.

How to get sharper CIs with fast rates, like Gaussian as the exp. decay?

Chernoff's inequality and Hoeffding's inequality

Chernoff's inequality is a good application of Markov's inequality with

$$\varphi(x) = e^{tx} \text{ implies } P(X \geq a) \leq e^{-ta}\mathrm{E}e^{tX} \text{ and minimize } t \text{ on } t > 0$$

Lemma 1.2 (Chernoff's inequality)

For a r.v. X with $\mathrm{E}e^{tX} < \infty$ on $t > 0$, $P(X \geq a) \leq \inf_{t>0}\left\{e^{-ta}\mathrm{E}e^{tX}\right\}$.

A sharper bound for **the sum of ind. r.vs** was attempted by Hoeffding.

Corollary 1.3 (Hoeffding's inequality for bounded r.v.)

Let $\{X_i\}_{i=1}^n$ be ind. r.vs satisfying the bound condition $a_i \leq X_i \leq b_i$. So,

$$P(|\sum_{i=1}^n (X_i - \mathrm{E}X_i)| \geq t) \leq 2e^{-2t^2/\sum_{i=1}^n (b_i - a_i)^2} \text{ for all } t \geq 0.$$

Sample $((X_i, Y_i))_{i=1}^n$, and define $Z_i := 1_{[f(X_i) \neq Y_i]}$ for classifier f. With p.a.l. $1 - \delta$,

$$\mathrm{P}[f(X) \neq Y] - \frac{1}{n}\sum_{i=1}^n 1\left[f\left(x_i\right) = y_i\right] = \mathrm{E}Z_1 - \frac{1}{n}\sum_{i=1}^n Z_i \leq \sqrt{\frac{1}{2n}\ln\left(\frac{1}{\delta}\right)}$$

Chernoff, H. (1952). A measure of asymptotic efficiency for tests of a hypothesis based on the sum of observations. AOMS.

Hoeffding, W. (1963). Probability inequalities for sums of bounded random variables. JASA, 58(301), 13-30.

Hoeffding's inequality works for confidence intervals

- For i.i.d. X_i's with $a \leq X_i \leq b$, Hoeffding inequality gives

$$P\left(\mu_0 \in [\bar{X}_n - \frac{b-a}{\sqrt{2}}\sqrt{\frac{1}{n}\log(\frac{2}{\delta})}], \bar{X}_n + \frac{b-a}{\sqrt{2}}\sqrt{\frac{1}{n}\log(\frac{2}{\delta})}]\right) \geq 1-\delta.$$

- Let us examine Bernoulli samples $\{X_i\}_{i=1}^n \overset{\text{i.i.d.}}{\sim} \text{Ber}\,(1/2)$, with $0 \leq X_i \leq 1$ and $\text{Var}X_i = 1/4$. Put $\delta = 0.05$, for any sample size n

$$P\left(\mu_0 \in [\bar{X}_n - \frac{1.36}{\sqrt{n}}, \bar{X}_n + \frac{1.36}{\sqrt{n}}]\right) \geq 95\%$$

which is sharp in the rate but not the constant in comparison with the asymptotic confidence interval, i.e.,

$$\lim_{n\to\infty} P\left(\mu_0 \in [\bar{X}_n - \frac{0.98}{\sqrt{n}}, \bar{X}_n + \frac{0.98}{\sqrt{n}}]\right) = 95\%.$$

Empirical comparison for Bernoulli r.vs.

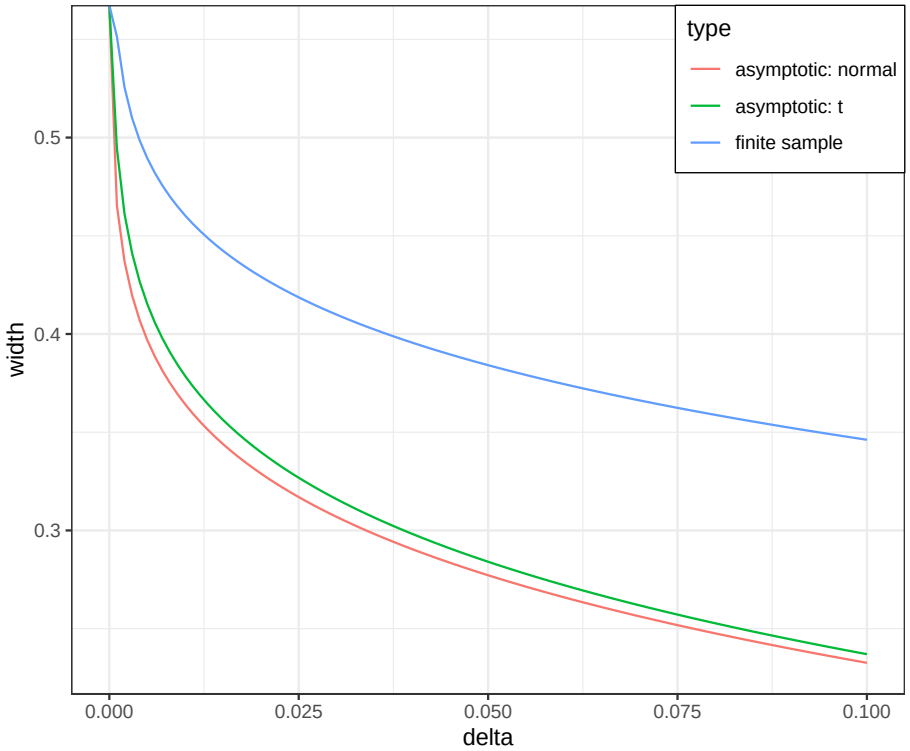


Figure: Width function for CIs.

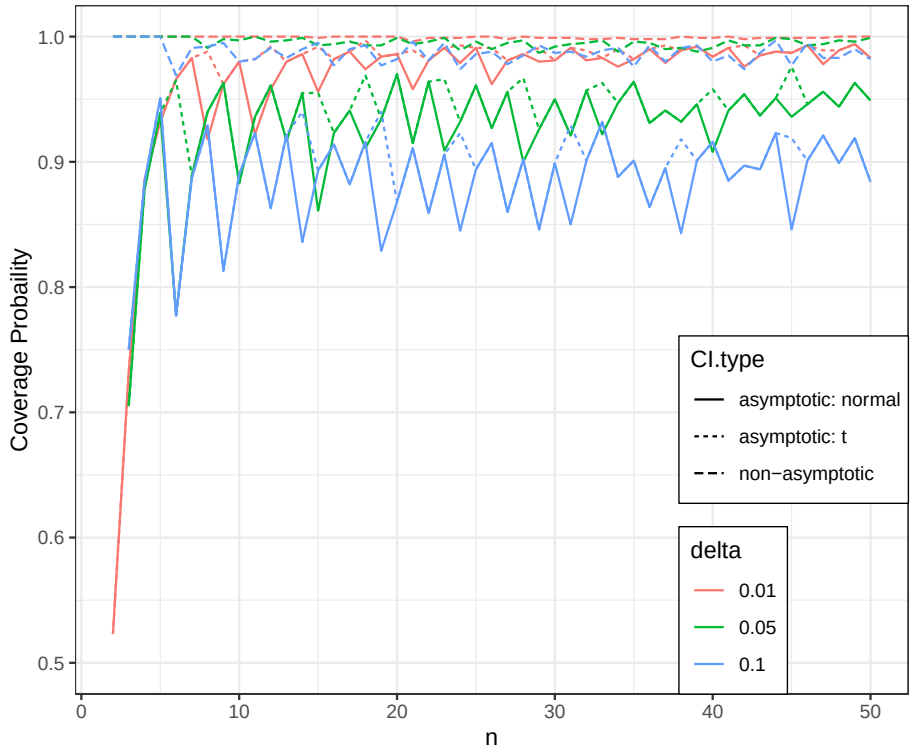
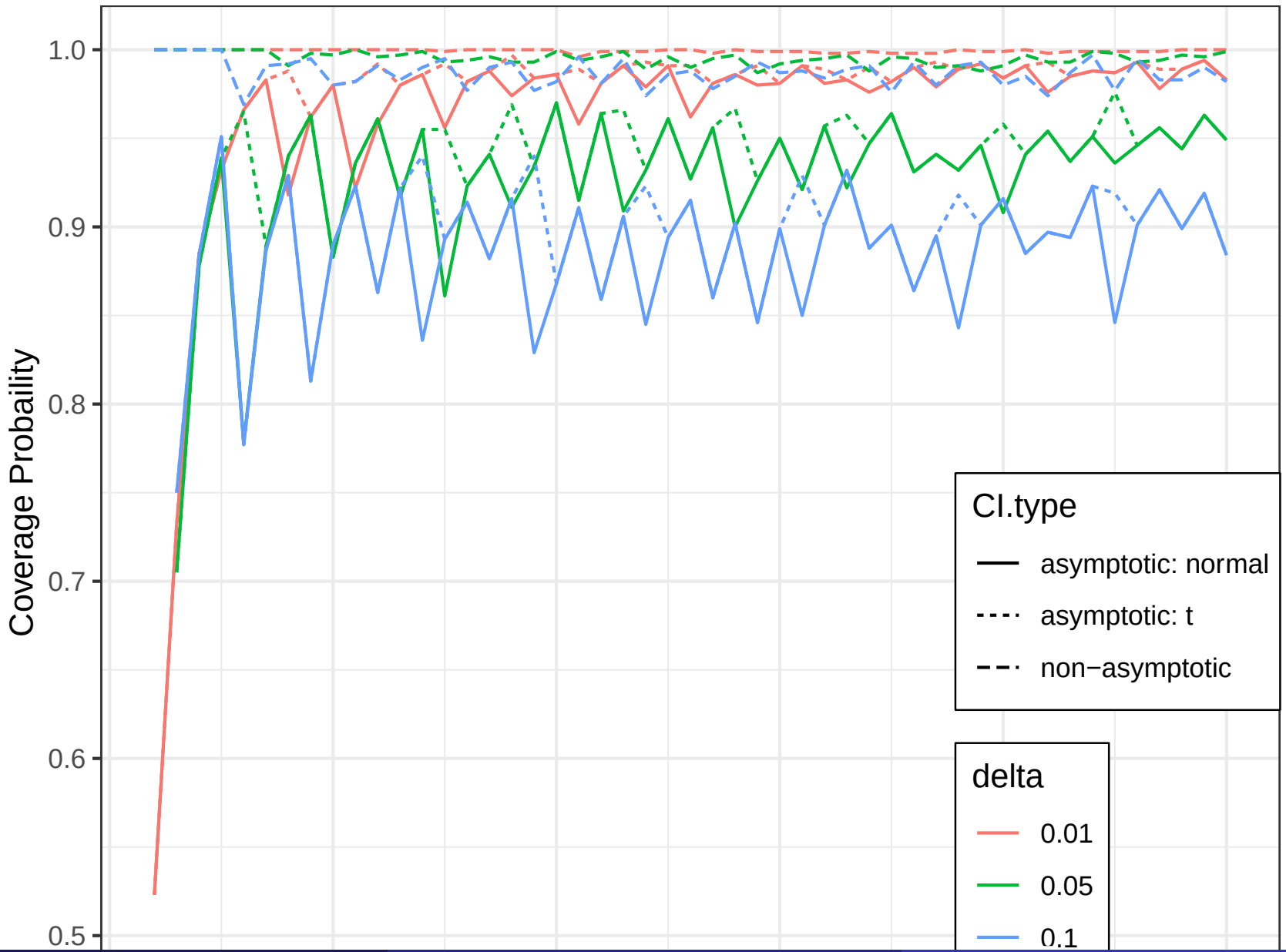


Figure: Coverage probabilities.

Left plot compares width of confidence intervals of **non-asymp(Hoeffding), student-t(with ture var.) and asymp methods** $n = 50$; right plot demonstrate their coverage probability versus (small) sample size and simulation were repeated 1000 times.

Two competing effects: Conservativeness vs Asymptotics

Empirical comparison for Bernoulli r.vs.



1.2 Non-asymptotic bounds in robust statistical learnings

The usually assumption that the data is i.i.d. is not plausible in robust statistical learning. We prefer to analysis the independent r.vs. $\{(X_i, Y_i)\}_{i=1}^n \in \mathbb{R}^p \times \mathbb{R}$

- Let $R_l^n(\theta) := \mathrm{E}[\frac{1}{n}\sum_{i=1}^n l(Y_i, X_i^\top \theta)]$ the risk function, the true parameter θ^* is defined as the minimizer for the empirical mean of the expected loss

$$\theta_n^* := \arg\min_{\theta\in\Theta} \mathrm{E}[\frac{1}{n}\sum_{i=1}^n l(Y_i, X_i^\top \theta)] = \arg\min_{\theta\in\Theta} R_l^n(\theta).$$

The sample-size dependent θ_n^* is different from i.i.d. version

$$\theta^* := \arg\min_{\theta\in\Theta} \mathrm{E}[l(Y_1, X_1^\top \theta)], \text{ which is free of } n.$$

- Given $l(\cdot,\cdot)$, the *empirical risk minimization* (ERM) is given by

$$\bar{\theta} := \arg\min_{\theta\in\Theta} \frac{1}{n}\sum_{i=1}^n l(Y_i, X_i^\top \theta) = \arg\min_{\theta\in\Theta} \hat{R}_l(\theta). \tag{1}$$

Usually, the ERM coincides with maximum likelihood estimation (MLE) and we want to find a candidate $\bar{\theta}$ that makes the MLE-based losses $\hat{R}_l(\theta)$ small.

- The **excess risk**

$$\mathcal{E}(\bar{\theta}, \theta^*) := R_l^n(\bar{\theta}) - R_l^n(\theta^*)$$

is a popular used measure of accuracy of $\bar{\theta}$ in the machine learning.

- We aim to a derive the sharp and optimal excess risk bounds

$$P\{R_l^n(\bar{\theta}) - R_l^n(\theta^*) \leq R_{n,p,\delta}(d, P_{\theta^*})\} \geq 1 - \delta,$$

for learning from general loss functions based on ERM without bound input assumptions, by the decomposition:

$$R_l(\bar{\theta}) - R_l^n(\theta_n^*) = \underbrace{R_l(\bar{\theta}) - \hat{R}_l(\bar{\theta})}_{\textit{Genaralization}} + \underbrace{\hat{R}_l(\bar{\theta}) - \hat{R}_l(\theta_n^*)}_{\textit{Optimaztion}} + \underbrace{\hat{R}_l(\theta_n^*) - R_l^n(\theta_n^*)}_{\textit{Concentration}} \tag{2}$$

$$\leq [R_l(\bar{\theta}) - \hat{R}_l^n(\bar{\theta})] + [\hat{R}_l^n(\theta_n^*) - R_l(\theta_n^*)], \text{ (By ERM)}$$

$$= \frac{1}{n}\sum_{i=1}^{n}\left(l(Y_i, X_i^\top \theta_n^*) - l(Y_i, X_i^\top \bar{\theta})\right) - \mathrm{E}[\frac{1}{n}\sum_{i=1}^{n}\left(l(Y_i, X_i^\top \theta_n^*) - l(Y_i, X_i^\top \bar{\theta})\right)]$$

$$\leq \sup_{\theta \in \Theta}\left[\frac{1}{n}\sum_{i=1}^{n}\left(l(Y_i, X_i^\top \theta_n^*) - l(Y_i, X_i^\top \theta)\right) - \mathrm{E}[\frac{1}{n}\sum_{i=1}^{n}\left(l(Y_i, X_i^\top \theta_n^*) - l(Y_i, X_i^\top \theta)\right)]\right]$$

2. Sub-Gaussian Distributions

In statistical ML, we want to MGF $\mathrm{E}e^{sX}$ have similar Gaussian MGF and tails.

$$P(|N(0, \mathrm{Var}(X))| \geq x) \leq e^{-x^2/[2\mathrm{Var}(X)]}\text{–}\textbf{Gaussian tail prob.}$$

Definition 2.1 (Sub-Gaussian distribution)

A zero-mean r.v. $X \in \mathbb{R}$ is sub-Gaussian with a **variance proxy** σ^2 if its MGF

$$\mathrm{E}e^{sX} \leq e^{\frac{\sigma^2 s^2}{2}}, \ \forall s \in \mathbb{R}. \ (\text{denoted } X \sim \mathrm{subG}\left(\sigma^2\right))$$

$$P\left(X \geq t\right) \leq \inf_{s>0} e^{-st}\mathrm{E}e^{sX} \leq \inf_{s>0} e^{-st+\frac{\sigma^2 s^2}{2}} \xlongequal{s=t/\sigma^2} e^{-\frac{t^2}{2\sigma^2}},$$

by Chernoff's inequality. This argument is called *Cramer-Chernoff method*.

- Let $Z_1, \ldots, Z_n$ be n independent centralized r.vs, and suppose there exists a convex function $g(t)$ and a domain D_0 containing $\{0\}$ such that

$$\mathrm{E}e^{t\sum_{i=1}^n Z_i} \leq e^{ng(t)}, \ \forall t \in D_0 \subset \mathbb{R}.$$

- Denote $g^*(s) = \sup_{t \in D_0}\{ts - g(t)\}$ as the *convex conjugate function* of g,

$$P(|\tfrac{1}{n}\sum_{i=1}^n Z_i| > s) \leq 2e^{-ng^*(s)}, \ \forall s > 0 \text{ by Chernoff's inequality.}$$

Example 2.2 (Normal distributions)

Consider the normal r.v. $X \sim N(\mu, \sigma^2)$ and $\mathrm{E}e^{sX} := e^{\frac{\sigma^2 s^2}{2}}$, $\forall s \in \mathbb{R}$, it is subG (σ^2) with the variance proxy $\sigma^2 = \mathrm{Var}(X)$.

Theorem 2.3 (Hoeffding's inequality)

Let $\{X_i\}_{i=1}^n$ be ind. r.vs satisfying the bound condition $a_i \le X_i \le b_i$. Then,

(a) **Hoeffding's lemma**: $\mathrm{E}e^{u\sum_{i=1}^n (X_i - \mathrm{E}X_i)} \le e^{\frac{u^2}{8}\sum_{i=1}^n (b_i - a_i)^2}$, $u \ge 0$;

(b) **Hoeffding's inequality**:
$P(|\sum_{i=1}^n (X_i - \mathrm{E}X_i)| \ge t) \le 2e^{-2t^2/\sum_{i=1}^n (b_i - a_i)^2}$, $t \ge 0$.

Hoeffding's inequality is proved by Cramer-Chernoff method.

Example 2.4 (Bounded r.vs)

Let $n = 1$. By Hoeffding's lemma, $\mathrm{E}e^{sX} \le e^{\frac{1}{8}s^2(b-a)^2}$ for $s > 0$ for the centralized bounded variable $X \in [a, b]$. So $X \sim$ subG $\left(\frac{1}{4}(b-a)^2\right)$.
For Bernoulli variable $X \in \{0, 1\}$, we have $X \sim$ subG $\left(\frac{1}{4}\right)$.

Properties of Sub-Gaussian Distributions

Variance inequality familiar [if $\{X_i\}_{i=1}^n$ are ind., $\mathrm{Var}(\sum_{i=1}^n X_i) = \sum_{i=1}^n \mathrm{Var}(X_i)$.]

Proposition 2.5 (Closed under i.d. sum and moment bounds)

(a). If $\{X_i\}_{i=1}^n$ are independent with $X_i \sim \mathrm{subG}\left(\sigma_i^2\right)$, then

$$\sum_{i=1}^n X_i \sim \mathrm{subG}\left(\sum_{i=1}^n \sigma_i^2\right).$$

(b). If $X \sim \mathrm{subG}(\sigma^2)$, $\mathrm{E}|X|^k \leq (2\sigma^2)^{k/2} k\Gamma(\frac{k}{2})$ and

$$k^{-1/2}(\mathrm{E}(|X|^k))^{1/k} \leq \sigma e^{1/e}, \ k \geq 2.$$

Proof: $\mathrm{E}e^{t(\sum_{i=1}^n X_i)} = \prod_{i=1}^n \mathrm{E}e^{tX_i} \leq \prod_{i=1}^n e^{\sigma_i^2 t^2/2} = e^{\sum_{i=1}^n \sigma_i^2 t^2/2}, \ \forall\, t \in \mathbb{R}$

- $\sigma^2 \geq \mathrm{Var}\, X$.The variance proxy σ^2 not only characterizes the speed of decay in the tail prob. but also is a upper bounds for the $\mathrm{Var}\, X$ as well.

  $$\frac{\sigma^2 s^2}{2} + o(s^2) = e^{\frac{\sigma^2 s^2}{2}} - 1 \geq \mathrm{E}e^{sX} - 1 = s\mathrm{E}X + \frac{s^2}{2}\mathrm{E}X^2 + \cdots = \frac{s^2}{2} \cdot \mathrm{Var}\, X + o\left(s^2\right)$$

  which implies by dividing s^2 on both sides and taking $s \to 0$.

Theorem 2.6 (Concentration for sub-Gaussian distributions)

Suppose $\{X_i\}_{i=1}^n$ *are independent with* $X_i \sim \text{subG}(\sigma_i)$,

$$\mathbb{P}(|\sum_{i=1}^n X_i| \geq t) \leq 2\exp\left\{-\frac{t^2}{2\sum_{i=1}^n \sigma_i^2}\right\}, \quad t \geq 0$$

Proposition 2.7 (Concentration for weighted E-F summation)

Let $\{Y_i\}_{i=1}^n$ *be a sequence of exponential family (EF) r.vs with density*

$$f(y_i;\theta_i) = c(y_i)\exp\{y_i\theta_i - b(\theta_i)\}.$$

We assume the **bounded variances condition***: there exist a compact set* Ω *and some constant* C_b *such that* $\sup_{\theta_i\in\Omega} \ddot{b}(\theta_i) \leq C_b^2$ *for all* i*. Let* $\mathbf{w} := (w_1, \cdots, w_n)^T \in \mathbb{R}^n$ *be a non-random vector and define* $S_n^w =: \sum_{i=1}^n w_i Y_i$.

(a) $S_n^w - \mathrm{E}S_n^w \sim \text{subG}(C_b^2\|\mathbf{w}\|_2^2)$ *and* $P\{|S_n^w - \mathrm{E}S_n^w| > t\} \leq 2e^{-t^2/(2C_b^2\|\mathbf{w}\|_2^2)}$;

(b) *Moments bound: Let* $C_n := S_n^w - \mathrm{E}S_n^w$*. For all integer* $k \geq 1$,
$\mathrm{E}|C_n|^k \leq k(2C_b^2)^{k/2}\Gamma(k/2)\,\|\mathbf{w}\|_2^k$.

Seven equivalent characterizations of sub-Gaussianity

Corollary 2.8 (Characterizations of sub-Gaussianity)

Let X be a r.v. in $\mathbb{R}$ with $\mathrm{E}X = 0$. Then, the following are equivalent for finite positive constants $\{K_i\}_{i=1}^7$.

(1) *The MGF of X:* $\mathrm{E}e^{sX} \leq e^{K_1^2 s^2}$ *for all* $s \in \mathbb{R}$;

(2) *The tail prob. of X:* $P\{|X| \geq t\} \leq 2e^{-t^2/K_2^2}$ *for all* $t \geq 0$;

(3) *The moments of X:* $(\mathrm{E}|X|^k)^{1/k} \leq K_3\sqrt{k}$ *for all integer* $k \geq 1$;

(4) *The exponential moment of X^2 :* $\mathrm{E}e^{X^2/K_4^2} \leq 2$;

(5) *The local MGF of X^2:* $\mathrm{E}e^{l^2X^2} \leq e^{K_5^2 l^2}$ *for all l in a local set* $|l| \leq \frac{1}{K_5}$.

(6) *There is a constant* $K_6 \geq 0$ *such that* $\mathrm{E}e^{\lambda X^2/K_6^2} \leq (1-\lambda)^{-1/2}\ \forall\ \lambda \in [0,1)$.

(7) *Maximal inequality. Let* $\{X_i\}_{i=1}^n$ *be i.i.d. copies of X, then* $\exists c > 0$ *s.t.*

$$\mathrm{E}\left[\max\{|X_1|, \ldots, |X_n|\}\right] \leq c\sqrt{\log n} \text{ for all } n \geq c.$$

Definition 2.9 (Optimal variance proxy, Buldygin and Kozachenko(2000))

The minimal σ^2 in subG MGF-def. is called the **optimal variance proxy**, i.e.

$$\sigma^2_{opt}(X) := \inf\left\{\sigma^2 \geq 0 : \mathrm{E}e^{tX} \leq e^{\sigma^2 t^2/2}, \forall\, t \in \mathbb{R}\right\}.$$

Definition 2.10 (Sub-Gaussian norm)

The sub-Gaussian norm of X is defined by: $\|X\|_{\psi_2} = \inf\{t > 0 : \ \mathrm{E}e^{X^2/t^2} \leq 2\}$.

- From Corollary 2.8(4), $\|X\|_{\psi_2}$ is the smallest K_4. An alternative def. of the sub-Gaussian norm is $\|X\|_{\psi_2} := \sup_{p\geq 1} p^{-1/2}(\mathrm{E}|X|^p)^{1/p}$(Vershynin,2010).
- a. No requirement for X having zero mean. b. X is subG iff $\|X\|_{\psi_2} < \infty$.
- In fact, if $\mathrm{E}e^{X^2/\|X\|^2_{\psi_2}} \leq 2$, then

$$P(|X| \geq t) = P(e^{X^2/\|X\|^2_{\psi_2}} \geq e^{t^2/\|X\|^2_{\psi_2}}) \leq \mathrm{E}e^{X^2/\|X\|^2_{\psi_2}}/e^{t^2/\|X\|^2_{\psi_2}} \leq 2e^{-t^2/\|X\|^2_{\psi_2}}.$$

Buldygin, V., & Kozachenko, Y. (2000). Metric characterization of random variables and random processes. AMS.

3. Sub-exponential Distributions

The requirement in def. of sub-Gaussian r.v. $\mathrm{E}e^{sX} \le e^{\frac{\sigma^2 s^2}{2}},\ \forall s \in \mathbb{R}$ is too strong.

Example 3.1 (MGF of exponential distributions)

Consider the exponential r.v. $X \sim \mathrm{Exp}(\mu)$ ($f(x) = \mu^{-1} e^{-x/\mu} \cdot 1(x > 0)$) with $\mathrm{E}X = \mu > 0$. The MGF of $X - \mu$ satisfies

$$\mathrm{E}e^{s(X-\mu)} = \frac{e^{-s\mu}}{1-s\mu} \le e^{2(s\mu/2)^2} < e^{s^2(2\mu)^2/2}, \forall\ |s| \le (2\mu)^{-1}$$

where the second last inequality is by $e^{-t}/\sqrt{1-2t} \le e^{2t^2}$ for $|t| \le 1/4$.

Definition 3.2 (Sub-exponential distributions)

A r.v. $X \in \mathbb{R}$ with $\mathrm{E}X = 0$ is sub-exponential with parameter λ if

$$\mathrm{E}e^{sX} \le e^{\frac{s^2\lambda^2}{2}} \quad \text{for all } |s| < 1/\lambda. \text{ (denoted } X \sim \mathrm{subE}(\lambda))$$

In Wainwright(2019), sub-E is defined by two positive parameters (λ, α):

$$\mathrm{E}e^{sX} \le e^{\frac{s^2\lambda^2}{2}} \quad \text{for all } |s| < 1/\alpha. \text{ (denoted } X \sim \mathrm{subE}(\lambda, \alpha))$$

Equivalent characterizations of sub-exponentiality

Corollary 3.3 (Characterizations of sub-exponentiality)

Let X be a r.v. in $\mathbb{R}$ with $\mathrm{E}X = 0$. Then the following properties are equivalent, where $\{K_i\}_{i=1}^6$ are positive constants.

(1) *The tails of X satisfy $P\{|X| \geq t\} \leq 2e^{-t/K_1}$ for all $t \geq 0$;*

(2) *The MGF of X satisfies $\mathrm{E}e^{lX} \leq e^{K_2^2 l^2}$ for all $|l| \leq \frac{1}{K_2}$;*

(3) *The moments of X satisfy $(\mathrm{E}|X|^p)^{1/p} \leq K_3 p$ for integer $p \geq 1$;*

(4) *The MGF of $|X|$ satisfies $\mathrm{E}e^{l|X|} \leq e^{K_4 l}$ for all $0 \leq l \leq \frac{1}{K_4}$;*

(5) *The MGF of $|X|$ is bounded at some point: $\mathrm{E}e^{|X|/K_5} \leq 2$;*

(6) **Bounded MGF of X in a compact set: $\mathrm{E}e^{tX} < \infty,\ \forall |t| < 1/K_6$.**

The (6) is the called *Cramer's condition* which is essential, it signifies that:

All r.vs. are sub-exponential if their MGF exist in a neighborhood of zero.

Proposition 3.4 (Concentration for weighted sub-exponential sums)

Let $\{X_i\}_{i=1}^n$ *be independent zero-mean* $\{\mathrm{subE}(\lambda_i, \alpha_i)\}_{i=1}^n$ *distributed. Define* $\alpha := \max_{1\le i\le n} \alpha_i > 0$, $\|\boldsymbol{\lambda}\|_2 := (\sum_{i=1}^n \lambda_i^2)^{1/2}$ *and* $\bar{\lambda} := (\frac{1}{n}\sum_{i=1}^n \lambda_i^2)^{1/2}$. *Then, 1.* Closed under i.d. sum $\sum_{i=1}^n X_i \sim \mathrm{subE}(\|\boldsymbol{\lambda}\|_2, \alpha)$ *and 2.* SubG+SubE decay

$$P(|\frac{1}{n}\sum_{i=1}^n X_i| \ge t) \le 2e^{-\frac{1}{2}(\frac{nt^2}{\bar{\lambda}^2} \wedge \frac{nt}{\alpha})} = \begin{cases} 2e^{-\frac{nt^2}{2\bar{\lambda}^2}}, & 0 \le t \le \frac{\bar{\lambda}^2}{\alpha} \\ 2e^{-\frac{nt}{2\alpha}}, & t > \frac{\bar{\lambda}^2}{\alpha} \end{cases}.$$

By considering two rate in $(\frac{nt^2}{\bar{\lambda}^2} \wedge \frac{nt}{\alpha})$ separately, we have

$$P(|\frac{1}{n}\sum_{i=1}^n X_i| \ge \bar{\lambda}\sqrt{\frac{2s}{n}} + \alpha \cdot \frac{2s}{n}) \le 2e^{-s}, \ \forall\ s \ge 0.$$

Example 3.5 (Laplace r.vs)

A r.v. X has a Laplace distribution ($\mathrm{Laplace}(\mu, b)$, $\mu \in \mathbb{R}, b > 0$) if its density function is $f(x) = e^{-\frac{|x-\mu|}{b}}/2b$. The $\mathrm{Laplace}(0, b)$ has representation

$$X - \mathrm{E}X := U - V \text{ where } U \text{ and } V \text{ are independent } \mathrm{Exp}(b) \text{ distributed.}$$

$\mathrm{Exp}(b) - b \sim \mathrm{subE}(b, 2b)$ by Exp 3.2; $U - V \sim \mathrm{subE}(\sqrt{2}b, 2b)$ by Cor. 3.4(c).

CI with relation to SLLN,CLT, and LIL

- Prop. 3.4 are non-asymptotically valid for any finite n. It also has asymptotical merit, which implies: *Strong Law of Large Numbers* (SLNN), *Central Limit Theorem* (CLT), and *Law of the Iterated Logarithm* (LIL).

Example 3.6 (SLNN)

1. The sample mean $\bar{X}_n = \frac{1}{n}\sum_{i=1}^n X_i$ for $\{X_i\}_{i=1}^n \overset{\text{i.i.d.}}{\sim} \text{subE}(\lambda)$ with mean μ.
   The B-C lemma $\sum_{n=1}^\infty P(|\bar{X}_n - \mu| > \varepsilon) \le \sum_{n=1}^\infty 2e^{-\frac{n}{2}(\frac{\varepsilon^2}{\lambda^2} \wedge \frac{\varepsilon}{\lambda})} < \infty$ implies
   $$\bar{X}_n \overset{a.s.}{\to} \mu.$$
2. Cor. 3.4(b) also implies the rate of convergence for $\bar{X}_n$ for all n with a high probability, i.e. $\bar{X}_n$ has **the non-asymptotic error bounds** by
   $$|\bar{X}_n - \mu| \le \sqrt{\frac{2\lambda^2 t}{n}} \vee \frac{2\lambda t}{n} = \begin{cases} \sqrt{\frac{2\lambda^2 t}{n}}, & n \ge 2t \text{ (slow global rate)} \\ \frac{2\lambda t}{n}, & n < 2t \text{ (fast local rate)} \end{cases}$$
   $\forall t > 0$ with the probability at least $1 - 2e^{-t}$.

Example 3.7 (CLT,LIL)

1. Applying Corollary 3.4(b), we have

$$P(|\sqrt{n}\bar{X}_n| \geq t) \leq 2\exp\left\{-\frac{1}{2}\left(\frac{t^2}{\lambda^2} \wedge \frac{t}{\lambda/\sqrt{n}}\right)\right\} = \begin{cases} 2e^{-ct^2/\lambda^2}, & t \leq \lambda\sqrt{n}; \\ 2e^{-t\sqrt{n}/\lambda}, & t > \lambda\sqrt{n}. \end{cases}$$

2. The CI indicates the phase transition about the tail behavior of $\sqrt{n}\bar{X}_n$:

   Small Deviation Regime. In the regime $t \leq \lambda\sqrt{n}$, we have a sub-Gaussian tail bound with variance proxy λ^2 as if the sum has the *normal distribution* with a constant variance as CLT

   Large Deviation Regime. In the regime $t \geq \lambda\sqrt{n}$, the sum has a heavier tail. The sub-exponential tail bound behaves as subE$(\lambda/\sqrt{n})$.

3. (LIL). Let $t/\sqrt{n} = R\sqrt{\log\log n}/\sqrt{n} \leq \lambda$ for $R > 0$. Corollary 3.4(b) claims

$$P(|\bar{X}_n| \geq \frac{R\sqrt{\log\log n}}{\sqrt{n}}) \leq 2e^{-t^2/2\|\mathbf{w}\|_2^2\lambda^2} = 2/(\log n)^{R^2/2\lambda^2}.$$

   Therefore, with probability $1 - 2/(\log n)^{R^2/2\lambda^2}$, $|\bar{X}_n| \leq \frac{R\sqrt{\log\log n}}{\sqrt{n}}$.

Jamieson, K. et al.(2014).lilucb: An optimal exploration algorithm for multi-armed bandits. In Conference on Learning Theory

Example 3.8 (Geometric distributions)

The *geometric distribution* $X \sim \mathrm{Geo}(q)$ for r.v. X is defined by:
$P(X = k) = (1-q)q^{k-1}, (q \in (0,1),\ k = 1, 2, \cdots)$.
Lemma 4.3 in Hillar and Wibisono (2013) shows $(\mathrm{E}|X|^k)^{1/k} < -2k/\log(1-q)$.
It follows from Minkowski's and Jensen's inequalities $(\mathrm{E}|Z|)^k \le \mathrm{E}|Z|^k$ for $k \ge 1$

$$(\mathrm{E}|X - \mathrm{E}X|^k)^{1/k} \le (\mathrm{E}|X|^k)^{1/k} + |\mathrm{E}X| \le 2(\mathrm{E}|X|^k)^{1/k} \le -4k/\log(1-q)$$

and Cor. 3.3(3) implies the *centralized* $\mathrm{Geo}(q)$ is sub-E with $K_3 = -4/\log(1-q)$.

Example 3.9 (Discrete Laplace r.vs)

A r.v. $X \sim \mathrm{DL}(q),\ q \in (0,1)$. obeys the discrete Laplace distribution if
$f_q(k) = P(X = k) = \frac{1-q}{1+q} q^{|k|},\ k \in \mathbb{Z} = \{0, \pm 1, \pm 2, \ldots\}$
The discrete Laplace r.v. is the difference of two i.i.d. $\mathrm{Geo}(q)$. The $\mathrm{Geo}(q)$ is sub-exponential, thus Cor. 3.4(a) implies that $\mathrm{DL}(q)$ is also sub-E. In differential privacy of network models, the noises in the degree sequence are assumed following $\mathrm{DL}(q)$, see Fan et al. (2020) and references therein.

Hillar, C., & Wibisono, A. (2013). Maximum entropy distributions on graphs. arXiv preprint arXiv:1301.3321.
Fan, Y., Zhang, H., & Yan, T. (2020). Asymptotic theory for differentially private generalized β-models with parameters increasing. Statistics and Its Interface, 13(3), 385-398.

Sub-exponential norm

Similar to the definition of sub-G norm, we define the sub-exponential norm.

Definition 3.10 (sub-exponential norm)

The sub-exponential norm of X is defined as

$$\|X\|_{\psi_1} = \inf\{t > 0 : \mathrm{E}\exp(|X|/t) \leq 2\}. \tag{3}$$

A second def. is $\|X\|_{\psi_1} := \sup_{p\geq 1} p^{-1}(\mathrm{E}|X|^p)^{1/p}$ as in Vershynin(2010).

Proposition 3.11 (Properties of sub-exponential norm)

If $\mathrm{E}\exp(|X|/\|X\|_{\psi_1}) \leq 2$, *then*

(a) *Tail bounds* $P(|X| > t) \leq 2e^{-t/\|X\|_{\psi_1}}$ *for all* $t \geq 0$;

(b) *Moment bounds* $\mathrm{E}|X|^k \leq 2\|X\|_{\psi_1}^k k!$ *for all integer* $k \geq 1$;

(c) *If* $\mathrm{E}X = 0$, *the MGF bounds* $\mathrm{E}e^{sX} \leq e^{(2\|X\|_{\psi_1})^2 s^2}$ *for all* $|s| < 1/(2\|X\|_{\psi_1})$,

$$X \sim \mathrm{subE}(2\|X\|_{\psi_1}).$$

Sub-exponential norm

Proposition 3.12 (Concentration for r.v. with sub-exponential sum)

Let $\{X_i\}_{i=1}^n$ be zero mean ind. sub-E distributed with $\|X_i\|_{\psi_1} \leq \infty$. Then,

$$P(|\sum_{i=1}^n X_i| \geq t) \leq 2\exp\{-\frac{1}{4}(\frac{t^2}{\sum_{i=1}^n 2\|X_i\|_{\psi_1}^2} \wedge \frac{t}{\max\limits_{1\leq i\leq n}\|X_i\|_{\psi_1}})\}, \ t \geq 0.$$

Proof.

If $\mathrm{E}\exp(|X|/\|X\|_{\psi_1}) \leq 2$, then $X \sim \mathrm{subE}(2\|X\|_{\psi_1})$ by using Proposition 3.11(c). The result follows by employing Corollary 3.4(b). □

Lemma 3.13 (Square and product of sub-Gaussian are sub-exponential)

(a). A r.v. X is sub-Gaussian iff X^2 is sub-exponential and $\|X^2\|_{\psi_1} = \|X\|_{\psi_2}^2$;
(b). Let X and Y be sub-Gaussian r.vs. Then XY is sub-exponential and

$$\|XY\|_{\psi_1} \leq \|X\|_{\psi_2}\|Y\|_{\psi_2}.$$

Proof of Lemma 3.13

The (a) is from def. To see (b), assume $\|X\|_{\psi_2} = \|Y\|_{\psi_2} = 1$, WLOG.
Apply Young's inequality $ab \le \left(a^2 + b^2\right)/2$ for all $a, b \in \mathbb{R}$, **twice**

$$\mathrm{E}e^{|XY|} \le \mathrm{E}e^{(X^2+Y^2)/2} = \mathrm{E}\left[e^{X^2/2}e^{Y^2/2}\right] \le \frac{1}{2}\mathrm{E}\left[e^{X^2} + e^{Y^2}\right] \le 2.$$

hence $\|XY\|_{\psi_1} \le 1 = \|X\|_{\psi_2}\|Y\|_{\psi_2}$.

Proposition 3.14

Let $X \sim \mathrm{subG}(\sigma^2)$, *then* $X^2 \sim \mathrm{subE}(8\sqrt{2}\sigma^2, 8\sigma^2)$.

Example 3.15 (Chi-squared r.vs)

- If $Z \sim N(0,1)$, then Z^2 is sub-exponential by
$$\mathrm{E}e^{\lambda(Z^2-1)} = \begin{cases} \frac{e^{-\lambda}}{\sqrt{1-2\lambda}} & \lambda < 1/2 \\ \infty & \lambda > 1/2 \end{cases}$$
- We have $\mathrm{E}e^{\lambda(Z^2-1)} \le e^{4\lambda^2/2}, |\lambda| < 1/4$, hence $Z^2 \sim \mathrm{subE}(2,4)$.

4. Sub-Gamma Distributions and Bernstein's Inequality

- **Bernstein-type inequalities** have more precise concentration, it originally is an extension of the Hoeffding's inequality with bounded assumption.
- As mentioned by Pollard(2015), the proof of Hoeffding's inequality with endpoints of the interval $[a, b]$ in Lemma 1.3 (with $n = 1$) crudely depends on the variance bound: (if $a \le X \le b$ without **variance information**)

$$\mathrm{Var}X = \mathrm{E}(X - \mathrm{E}X)^2 \le \mathrm{E}[X - (b-a)/2]^2 \le [(b-a)/2]^2,$$

Corollary 4.1 (Bernstein's inequality with the bounded condition)

Let $\{X_i\}_{i=1}^n$ be centralized ind. r.vs. such that $|X_i| \le M$ a.s. for all i. Then,

$$P(|S_n| \ge t) \le 2e^{-\frac{t^2/2}{\sum_{i=1}^n \mathrm{Var}X_i + Mt/3}}, \quad P\{|S_n| \ge (2t\sum_{i=1}^n \mathrm{Var}X_i)^{1/2} + Mt/3\} \le 2e^{-t}.$$

$$\text{When } \mathrm{Var}X \ll c^2. \text{ Hoeffding}: \quad P(|\bar{X} - \mu| \le \sqrt{\frac{2c^2 \log(2/\delta)}{n}}) \ge 1 - \delta;$$

$$\text{Bernstein is sharper}: \quad P(|\bar{X} - \mu| \le \frac{c}{3n}\log(2/\delta) + \sqrt{\frac{2(\mathrm{Var}\,X)\log(2/\delta)}{n}}) \ge 1 - \delta.$$

Pollard, D. (2015). Mini Empirical. http://www.stat.yale.edu/pollard/Books/Mini.

Bernstein's CI is sharper

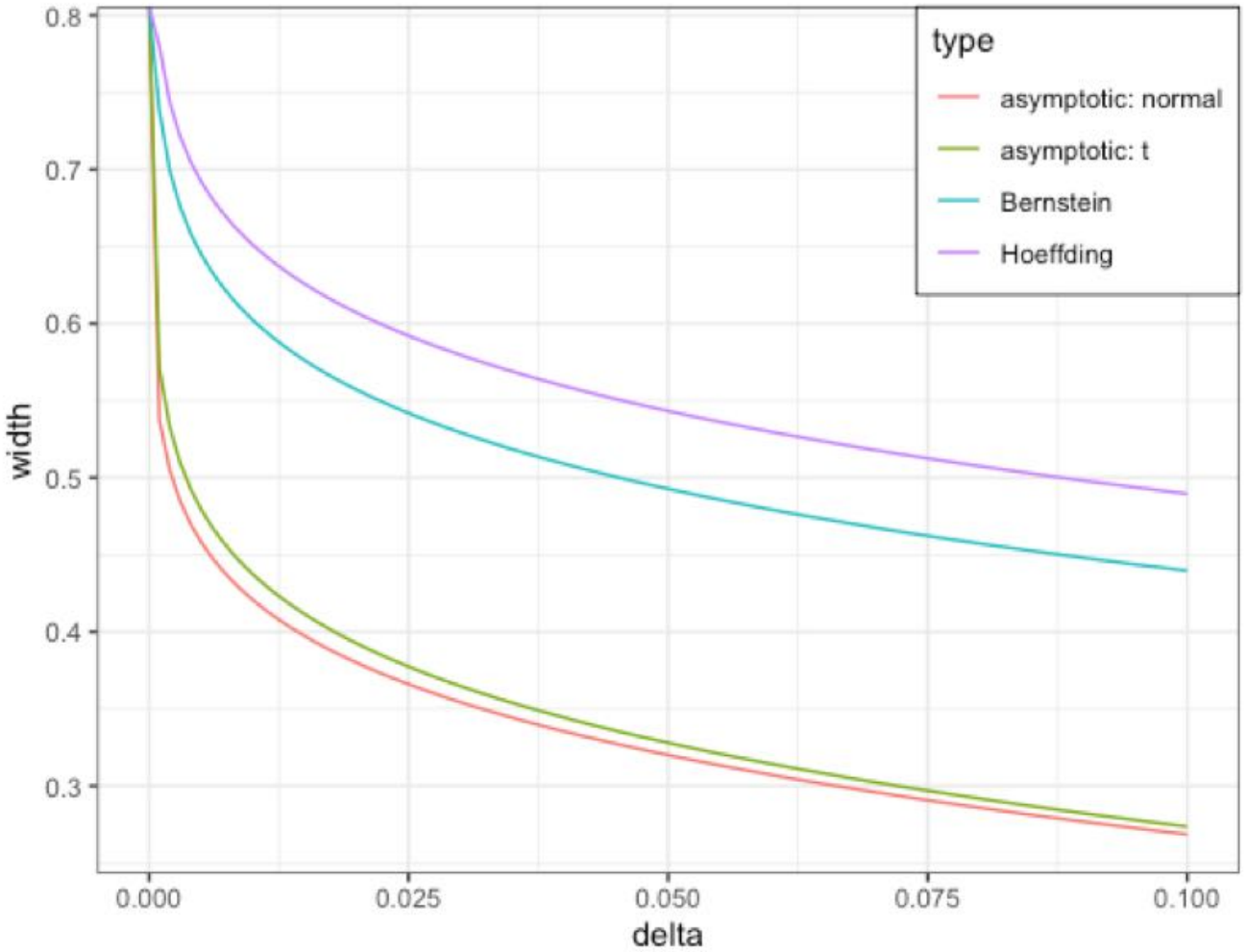


Figure: Width function for CIs. $\{X_i\}_{i=1}^n \overset{\text{i.i.d.}}{\sim} U[-1,1])$, $n = 50$.

Sub-Gamma Distributions

Example 4.2 (Gamma r.vs, page28 of Boucheron et al. (2013))

The Gamma distribution $\Gamma(a, b)$ with density $f(x) = \frac{x^{a-1}e^{-x/b}}{\Gamma(a)b^a}$, $x \geq 0$. We have $\mathrm{E}X = ab$ and $\mathrm{Var}\, X = ab^2$ for $X \sim \Gamma(a, b)$.

$$\log(\mathrm{E}e^{s(X-\mathrm{E}X)}) \leq s^2ab^2/[2(1-bs)] = s^2\mathrm{Var}\, X/[2(1-\frac{\mathrm{Var}\, X}{\mathrm{E}X}s)], \quad \forall\ 0 < s < b^{-1}.$$

Definition 4.3 (Sub-Gamma r.v.)

A centralized r.v. X is *sub-Gamma* with the *variance factor* $\upsilon > 0$ and the *scale parameter* $c > 0$ (denoted by $X \sim \mathrm{sub}\Gamma(\upsilon, c)$) if

$$\log(\mathrm{E}e^{sX}) \leq s^2\upsilon/[2(1-c|s|)], \quad \forall\ 0 < |s| < c^{-1}. \tag{4}$$

Example 4.4 (Sub-exponential r.vs)

The sub-exponential distribution with positive support implies the sub-Gamma condition: $\log(\mathrm{E}e^{sX}) \leq \frac{s^2\lambda^2}{2} \leq \frac{s^2\lambda^2}{2(1-\lambda|s|)}$, $\forall\ |s| < \frac{1}{\lambda}$. This shows that $X \sim \mathrm{subE}(\lambda)$ implies $X \sim \mathrm{sub}\Gamma(\lambda^2, \lambda)$.

Properties of Sub-Gamma Distributions

The sub-Gamma condition leads to the useful tail bounds and moment bounds.

Proposition 4.5 (Concentration for sub-Gamma sum)

Let $\{X_i\}_{i=1}^n$ *be independent* $\{\mathrm{subГ}(v_i, c_i)\}_{i=1}^n$ *distributed with zero mean. Then*

(a) *Closed under i.d. sum:*

$$S_n := \sum_{i=1}^n X_i \sim \mathrm{subГ}(\sum_{i=1}^n v_i, c),$$

where $c = \max_{1\le i\le n} c_i$;

(b) *Tail bounds:*

$$P(|S_n| \ge t) \le 2\exp\left(-\frac{t^2/2}{\sum_{i=1}^n v_i + ct}\right); P\{|S_n| > (2t\sum_{i=1}^n v_i)^{1/2} + ct\} \le 2e^{-t};$$

(c) *If* $X \sim \mathrm{subГ}(v, c)$, *the moments bounds satisfy for any integer* $k \ge 1$:

$$\mathrm{E}X^k \le k2^{k-2}[2(\sqrt{2v})^k \Gamma(\tfrac{k}{2}) + c(\sqrt{2v})^{k-1}\Gamma(\tfrac{k+1}{2}) + 3c^k\Gamma(k)].$$

Bernstein's inequality as Example of sub-Gamma CI

- In some statistical settings, one can not assume bounded r.vs.
- Bernstein's inequality for the sum of independent r.vs allows us to estimate the tail probability by a weaker version of an exponential moment on **the growth of the k-moment** without the boundedness.

Corollary 4.6 (Bernstein's inequality with the growth of moment condition)

If the centred ind. $\{X_i\}_{i=1}^n$ *satisfy* **the growth of moments condition**

$$\mathrm{E}|X_i|^k \leq 2^{-1} v_i^2 \kappa_i^{k-2} k!, \ (i = 1, 2, \cdots, n), \ \textit{for all } k \geq 2$$

where $\{\kappa_i\}_{i=1}^n, \{v_i\}_{i=1}^n$ *are constants independent of* k. *Let* $\nu_n^2 = \sum_{i=1}^n v_i^2$ *(the fluctuation of sums) and* $\kappa = \max_{1 \leq i \leq n} \kappa_i$ *(max scale). Then, we have* $X_i \sim \mathrm{subГ}(v_i, \kappa_i)$ *and for* $t > 0$

$$P\left(|S_n| \geq t\right) \leq 2e^{-\frac{t^2}{2\nu_n^2 + 2\kappa t}}, \quad P(|S_n| \geq \sqrt{2\nu_n^2 t} + \kappa t) \leq 2e^{-t}. \tag{5}$$

Concentration of exponential family

Example 4.7 (Normal r.v. $X_i \sim N(\mu, \sigma^2)$)

$$\mathrm{E}X_i^{2k-1} = 0;\ \mathrm{E}|X_i|^{2k} = \sigma^{2k}(2k-1)(2k-3)\cdots 3\cdot 1 \le 2^{-1}(2\sigma^2)\sigma^{2k-2}(2k)!,$$

which satisfies the growth of moments condition with $v_i^2 = 2\sigma^2, \kappa_i = \sigma^2$.

Theorem 4.8 (Sub-E concentration of exponential family)

Let $\{Y_i\}_{i=1}^n$ be a sequence of independent r.vs with their densities $\{f(y_i;\theta_i)\}_{i=1}^n$ belong to canonical exponential family on the natural parameter space $\theta_i \in \Theta$. Given non-random weights $\{w_i\}_{i=1}^n$ with $w = \max_{1\le i\le n}|w_i| > 0$, then

$$P(|\sum_{i=1}^n w_i(Y_i - \mathrm{E}Y_i)| \ge t) \le 2\exp(-\frac{t^2}{4w^2\sum_{i=1}^n C_{\theta_i} + 2w\max\limits_{1\le i\le n} C_{\theta_i} t}),$$

where $C_{\theta_i} := \inf_{0<r\le \mathrm{r}(\Theta)} r^{-1}[\mathrm{E}_{\theta_i} e^{r|Y_i - \dot{b}(\theta_i)|}]$.

Thm. 4.8 has no compact space assumption. If we impose the compact space assumption in Prop. 2.7, it leads to the sub-G concentration as presented before.

Concentration for functions of random vectors

McDiarmid's inequality (also called bounded difference inequality) is a CI for a multivariate function of random sequence $\{X_i\}_{i=1}^n$, says $f(X_1, ..., X_n)$.

Lemma 5.1 (McDiarmid's inequality)

Suppose $X_1, \cdots, X_n$ are independent r.vs all taking values in the set $\mathcal{X}$, and assume $f : \mathcal{X}^n \to \mathbb{R}$ satisfies the **bounded difference condition***(BDC)*

$$\sup_{x_1,\cdots,x_n,x_k' \in \mathcal{X}} |f(x_1, \cdots, x_n) - f(x_1, \cdots, x_{k-1}, x_k', x_{k+1}, \cdots, x_n)| \leq c_k.$$

Then, $P\left(|f(X_1, \cdots, X_n) - \mathrm{E}\left\{f(X_1, \cdots, X_n)\right\}| \geq t\right) \leq 2e^{-2t^2/\sum_{i=1}^n c_i^2} \quad \forall t > 0.$

- The requirement of the BDC is by replacing X_j by X_j' meanwhile maintaining the others fixed in $f(X_1, ..., X_n)$. The BDC is a Lipschitz property of f with respect to the Hamming distance.
- The set $\mathcal{B}(\mathbf{c})$ is the set of functions $f : \mathcal{X}^n \to \mathbb{R}$ such that, for any $x = (x_1, \ldots, x_n)$ and $y = (y_1, \ldots, y_n)$ in $\mathcal{X}^n$,

$$|f(x) - f(y)| \leq \sum_{i=1}^n c_i 1_{\{x_i \neq y_i\}}.$$

Two typical examples are the concentration for **U-statistics** (a dependent summation) and **the suprema of empirical processes**.

Example 5.2 (U-statistics)

Let $\{X_i\}_{i=1}^n$ be i.i.d. r.vs and $g : \mathbb{R}^2 \to \mathbb{R}$ be the bounded and symmetric function. Define a *U-statistic of order 2* as

$$U_n = \binom{n}{2}^{-1} \sum_{i<j} g(X_i, X_j) := f(x_1, ..., x_n).$$

Check bounded difference condition:

$$\begin{aligned} &|f(x_1, \ldots, x_{k-1}, x_k, x_{k+1}, \ldots x_n) - f(x_1, \ldots, x_{k-1}, x_k', x_{k+1}, \ldots x_n)| \\ &= \frac{1}{\binom{n}{2}} \Big| \sum_{j=1, j\neq k}^{n} [g(x_k, x_j) - g(x_k', x_j)] \Big| \leq \frac{2 \cdot 2(n-1)\|g\|_\infty}{n(n-1)} = \frac{4\|g\|_\infty}{n}. \end{aligned}$$

So we have

$$P(|U_n - \mathrm{E}U_n| > t) \leq 2e^{-nt^2/8\|g\|_\infty^2}.$$

The bounded difference inequality is of particular interest when

$$f(X) \in \textbf{supremum of bounded empirical processes (EP).}$$

Example 5.3 (The supremum of bounded EP)

Let $X = (X_1, \ldots, X_n)$ denote a set of independent $\mathcal{X}$-valued random vectors. Given $f \in \mathcal{F}$, WLOG, assume that

$$\forall f \in \mathcal{F}, \quad \mathrm{E}\left[f(X_i)\right] = 0, \quad \text{and} \quad f(X_i) \in [a_i, b_i]$$

- For any $x = (x_1, \ldots, x_n) \in \mathcal{X}^n$, define $f(x) = \sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^n f(x_i)$,

$$|\sup_{f\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^n f(x_i) - \sup_{f\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^n f(y_i)| \le \sup_{f\in\mathcal{F}} |\frac{1}{n}\sum_{i=1}^n (f(x_i) - f(y_i))| \le \sum_{i=1}^n \frac{b_i - a_i}{n} 1_{\{x_i \neq y_i\}}.$$

It is clear that $f \in \mathcal{B}(\mathbf{c})$, with $c_i = (b_i - a_i)/n$.

- Therefore, the BDC implies $\forall \sqrt{n}u \ge 0$

$$P(\sup_{f\in\mathcal{F}} \frac{1}{\sqrt{n}} \sum_{i=1}^n f(X_i) \le \mathrm{E}[\sup_{f\in\mathcal{F}} \frac{1}{\sqrt{n}} \sum_{i=1}^n f(X_i)] + \sqrt{n}u) \ge 1 - e^{-\frac{2(\sqrt{n}u)^2}{[\frac{1}{n}\sum_{i=1}^n (b_i - a_i)^2]}}.$$

Lemma 5.4 (Symmetrization Theorem)

Let $\varepsilon_1, ..., \varepsilon_n$ be a Rademacher sequence with uniform distribution on $\{-1, 1\}$, independent of $X_1, ..., X_n \in \mathcal{X}$ and $f \in \mathcal{F}$. Then we have

$$\mathrm{E}\left[\sup_{f\in\mathcal{F}}\left|\sum_{i=1}^{n}\left[f(X_i)-\mathrm{E}\left\{f(X_i)\right\}\right]\right|\right] \leq 2\mathrm{E}\left[\mathrm{E}_\epsilon\left\{\sup_{f\in\mathcal{F}}\left|\sum_{i=1}^{n}\epsilon_i f(X_i)\right|\right\}\right].$$

where $\mathrm{E}[\cdot]$ refers to the expectation w.r.t. $X_1, ..., X_n$ and $\mathrm{E}_\epsilon\{\cdot\}$ w.r.t. $\epsilon_1, ..., \epsilon_n$.

Example 5.5 (Moment bounds for suprema of EP)

Using Symmetrization Theorem,

$$\mathrm{E}[\sup_{f\in\mathcal{F}}\frac{1}{\sqrt{n}}\sum_{i=1}^{n}f(X_i)] \leq \mathrm{E}[\sup_{f\in\mathcal{F}}\frac{2}{\sqrt{n}}\sum_{i=1}^{n}\epsilon_i f(X_i)]$$

can easily be evaluated when $\mathcal{F}$ is a set linear functionals.
Let $r > 0, \|\cdot\|$ denote the Euclidean norm on $\mathbb{R}^p$ and $r\mathbf{B} = \{\mathbf{a} \in \mathbb{R}^p : \|\mathbf{a}\| \leqslant r\}$

$$\mathcal{F} = \left\{f : \mathbb{R}^p \to \mathbb{R} : \exists\, \mathbf{a} \in r\mathbf{B}, f(\mathbf{x}) = \mathbf{a}^T\mathbf{x}\right\}$$

$$\mathrm{E}[\sup_{f\in\mathcal{F}}\frac{1}{\sqrt{n}}\sum_{i=1}^{n}f(X_i)]\leq \mathrm{E}[\sup_{\mathbf{a}\in r\mathbf{B}}\frac{2}{\sqrt{n}}\sum_{i=1}^{n}\epsilon_i\mathbf{a}^T X_i]=\mathrm{E}[\sup_{\mathbf{a}\in r\mathbf{B}}\mathbf{a}^T(\frac{2}{\sqrt{n}}\sum_{i=1}^{n}\epsilon_i X_i)]$$

$$\text{(Cauchy's inequality)}\ \leq r\mathrm{E}\|\frac{2}{\sqrt{n}}\sum_{i=1}^{n}\epsilon_i X_i\|\leq r(\mathrm{E}\|\frac{2}{\sqrt{n}}\sum_{i=1}^{n}\epsilon_i X_i\|^2)^{1/2}$$

$$=\frac{2r}{\sqrt{n}}(\sum_{1\leqslant i,j\leqslant n}\mathrm{E}\{\mathrm{E}\left[\epsilon_i\epsilon_j X_i^T X_j\right]|\mathbf{X}\})^{1/2}$$

$$(\{\varepsilon_i\}_{i=1}^n\text{ cond. i.i.d.})\ =\frac{2r}{\sqrt{n}}(\sum_{1\leqslant i,j\leqslant n}\mathrm{E}\{\epsilon_i\epsilon_j\mathrm{E}[X_i^T X_j]\})^{1/2}=2r(\frac{1}{n}\sum_{i=1}^{n}\mathrm{E}[X_i^T X_i])^{1/2}$$

Finally, combine the McDiarmid's, with prob. at least $1-e^{-2M^2/[\frac{1}{n}\sum_{i=1}^{n}(b_i-a_i)^2]}$,

$$\sup_{f\in F}\frac{1}{n}\sum_{i=1}^{n}f(X_i)\leq \mathrm{E}[\sup_{f\in F}\frac{1}{n}\sum_{i=1}^{n}f(X_i)]+\frac{M}{\sqrt{n}}\leq\frac{1}{\sqrt{n}}[2r(\frac{1}{n}\sum_{i=1}^{n}\mathrm{E}[X_i^T X_i])^{1/2}+M].$$

Remark:$\mathrm{E}[\sup_{f\in\mathcal{F}}\frac{1}{\sqrt{n}}\sum_{i=1}^{n}f(X_i)]$ is bounded by the *uniform entropy integral* evaluated by VC dimension of the general $\mathcal{F}$, see Theorem 3.5.4 in Giné and Nickl(2015).

Giné, E., Nickl, R. (2015). Mathematical foundations of infinite-dimensional statistical models. Cambridge University Press.

In the analyses of high-dim. regression by EP, beyond McDiarmid's inequality, researches often resort to CIs of Lipschitz functions for strongly log-concave r.v.

Lemma 5.6 (Theorem 2.26, Wainwright(2019))

Let $\mathbf{N} \sim N(\mathbf{0}, \mathrm{I}_p)$. *Let* $f : \mathbb{R}^n \to \mathbb{R}$ *be L-Lipschitz with respect to (w.r.t.) the Euclidean norm:* $|f(\mathbf{a}) - f(\mathbf{b})| \leq L\|\mathbf{a} - \mathbf{b}\|_2$ *for any* $\mathbf{a}, \mathbf{b} \in \mathbb{R}^n$. *Then,*

$$P\left(|f(\mathbf{N}) - \mathrm{E}f(\mathbf{N})| \geq t\right) \leq 2e^{-t^2/(2L^2)}, \ \forall\ t > 0.$$

A function $\psi(\mathbf{x}) : \mathbb{R}^n \to \mathbb{R}$ is γ-*strongly concave* if there is some $\gamma > 0$ s.t.

$$\lambda\psi(\mathbf{x}) + (1-\lambda)\psi(\mathbf{y}) - \psi(\lambda\mathbf{x} + (1-\lambda)\mathbf{y}) \leq \frac{\gamma}{2}\lambda(1-\lambda)\|\mathbf{x}-\mathbf{y}\|_2^2, \ \forall\, \lambda \in [0,1] \text{ and } \mathbf{x}, \mathbf{y} \in \mathbb{R}^n,$$

A continuous probability density $f(\mathbf{x})$ and the corresponding r.v. is strongly log-concave if $f(\mathbf{x})$ is a strongly log-concave function. (**Hard to check!**)

Lemma 5.7 (Theorem 3.16, Wainwright(2019))

Let $\mathbb{P}$ *be any* $\gamma-$*strongly log-concave distribution on* $\mathbb{R}^n$ *with parameter* $\gamma > 0$. *Then for any function* $f : \mathbb{R}^n \to \mathbb{R}$ *that is L-Lipschitz w.r.t. the Euclidean norm,*

$$P[f(X) - \mathrm{E}f(X) \geq t] \leq e^{-\frac{\gamma t^2}{4L^2}} \ \text{for } X \sim \mathbb{P} \text{ and } t \geq 0.$$

Concentration for the suprema of unbounded empirical processes

Some times $\gamma-$strongly log-concave distribution is hard to verified.

- Let $Z = (Z_1, \ldots, Z_n)$ be a vector of **independent r.vs** with values in a space $\mathcal{Z}$, define Z' as an independent copy of Z.
- Consider a function $f : \mathcal{Z}^n \to \mathbb{R}$. It is of interest to study the CI for $f(Z)$.
- For $w \in \mathcal{Z}$ and $k \in \{1, \ldots, n\}$ define the **substitution operator** $S_w^k : \mathcal{Z}^n \to \mathcal{Z}^n$ by

$$S_w^k z = (z_1, \ldots, z_{k-1}, w, z_{k+1}, \ldots, z_n)$$

and the centered conditional version of f as the r.v.

$$\begin{aligned} Y_{f,Z_k}(z) &\equiv f\left(z_1, \ldots, z_{k-1}, Z_k, z_{k+1}, \ldots, z_n\right) - \mathbb{E}\left[f\left(z_1, \ldots, z_{k-1}, Z_k', z_{k+1}, \ldots, z_n\right)\right] \\ &= f\left(S_{Z_k}^k z\right) - \mathbb{E}[f(S_{Z_k'}^k z)] = \mathbb{E}[f\left(S_{Z_k}^k z\right) - f(S_{Z_k'}^k z) | Z_k] \end{aligned} \tag{6}$$

where $Y_{f,Z_k}(z)$ can be viewed as **random-valued functions** $z \in \mathcal{Z}^n \mapsto Y_{f,Z_k}(z)$. The McDiarmid's inequality with **unbounded difference conditions** is helpful to

Derive CI for the suprema of unbounded EP.

For any scalar random variable Z, the sub-G and sub-E norms $\|\cdot\|_{\psi_1}$ and $\|\cdot\|_{\psi_2}$ are defined as $\|Z\|_{\psi_1} = \sup_{p\geq 1} \|Z\|_p/p$ and $\|Z\|_{\psi_2} = \sup_{p\geq 1} \|Z\|_p/\sqrt{p}$.

Theorem 5.8 (Theorems 3.1 and 3.2 in Maurer and Pontil(2021))

Let K_X be the optimal sub-Gaussian parameter defined as $K_X := \inf\left\{K > 0 : \mathrm{E}e^{sX} \leq e^{K^2\|X\|_{\psi_2}^2 s^2}, \forall\, s \in \mathbb{R}\right\}$. If $\{Y_{f,X_k}(x)\}_{i=1}^n$ have finite sub-Gaussian norm, for $t > 0$ we have

$$\delta(X,t) := \mathrm{P}\left\{f(X) - \mathrm{E}f(X) > t\right\} \leq \exp\left(\frac{-t^2}{16K_X^2 \sup\limits_{x\in\mathcal{X}^n} \sum_k \|Y_{f,X_k}(x)\|_{\psi_2}^2}\right). \quad (7)$$

If $\{Y_{f,Z_k}(z)\}_{i=1}^n$ have finite sub-exponential norm, for $t > 0$ one has

$$\delta(X,t) \leq \exp\left(\frac{-t^2}{4e^2 \sup\limits_{x\in\mathcal{X}^n} \sum_k \|Y_{f,X_k}(x)\|_{\psi_2}^2 + 2e \max\limits_{1\leq k\leq n} \sup\limits_{x\in\mathcal{X}^n} \|Y_{f,X_k}(x)\|_{\psi_1} t}\right).$$

For example, if $f(x) = \sum_{i=1}^n x_i$ then $Y_{f,X_k}(x) = X_k - \mathrm{E}X_k$ is independent of x. So

$$\sup_{x\in\mathcal{X}^n} \sum_k \|Y_{f,X_k}(x)\|_{\psi_2}^2 = \sum_k \|X_k - \mathrm{E}[X_k]\|_{\psi_2}^2.$$

Maurer, A., & Pontil, M. (2021). Some Hoeffding-and Bernstein-type Concentration Inequalities. arXiv preprint.

Thanks